On Inter-Rater Reliability for Crowdsourced QoE

Tobias Hoßfeld*, Michael Seufert*, Babak Naderi†
*University of Würzburg, Chair of Communication Networks, Würzburg, Germany
†Technische Universität Berlin, Quality and Usability Lab, Berlin, Germany
E-mail: tobias.hossfeld@uni-wuerzburg.de, michael.seufert@uni-wuerzburg.de, babak.naderi@tu-berlin.de

***Abstract*—Crowdsourcing offers a faster, cheaper, and more scalable approach than the traditional laboratory quality assessment tests. However, participants perform the test in their own working environment, using their own hardware and without direct supervision of a test moderator, leading to different types of biases on the ratings. In this paper, we compare several reliability metrics that are commonly applied to the subjective ratings in terms of their sensitivity to identify typical issues of crowdsourced media quality tests. Following the subject bias theory, we simulate the ratings of different user groups with different bias and various magnitudes of uncertainty, while also considering the presence of unreliable raters. We apply traditional reliability metrics on the ratings and compare their sensitivity in identifying the severity of the raters' biases and uncertainties. Our results show that the average Spearman's rank correlation coefficient between raters can serve as a strong indicator for issues with the crowdsourcing study. This means that scoring too low for this metric should encourage researchers to revisit their study design in order to eventually improve the reliability of results from crowdsourcing-based quality studies.**

I. Introduction

Traditionally, the media's perceived Quality of Experience (QoE) [1] is assessed by performing subjective tests in controlled laboratory environments following the relevant recommendations. During the passive tests, representative group of participants, are invited to the laboratory and exposed to different test conditions by listening and/or watching a set of stimuli. The Absolute Category Rating (ACR) test procedure is the most common test method, in which participants are asked to rate their opinion about the overall quality of each stimulus on a predefined 5-point discrete scale, ranging from bad (1) to excellent (5). The test is executed on well-defined devices in a controlled laboratory environment to avoid that participants' ratings are influenced by unwanted additional factors. This leads to a high reliability and validity in the measurement, but sacrifices some realism, as the used devices and the test environments typically do not reflect normal usage situations.

Meanwhile, crowdsourcing offers a faster, cheaper, and more scalable approach in subjective testing [2]. There, test participants are workers from a crowdsourcing platform, who perform the QoE test in their working environment using their hardware, usually in exchange for monetary reward. The uncontrolled environment, heterogeneous and uncalibrated devices may lead to positive and negative biases in the ratings. For instance, in speech QoE tests, the environment noise might mask the degradation under study leading to either higher ratings, e.g., when the degradation mixes with background noise, or lower ratings, e.g., when the degradation affects the loudness, compared to ratings collected in a controlled laboratory environment [3]. To minimize the effect of such unwanted factors, the eligibility of test participants (e.g., normal hearing ability in speech quality test, or color vision deficiencies for video quality test), as well as the suitability of working environment (e.g., quiet environment [4], or bad lighting) and the devices (headset, or monitor with proper size and setting) should be examined before the test [5].

Besides, due to the absence of a test moderator, some participants might not fully follow the instruction, be interrupted, or rush the process to increase their hourly payout, which can lead to random ratings [2]. Those cases can partially be captured by gold standard [6] and trapping questions [7]. During the post-processing of submission, sessions in which participant failed in any of the integrated tests should be removed. It is also recommended to remove submissions, which show specific patterns in the ratings, or which are flagged by outlier detection methods. Although previous works showed that applying the best practices produce highly reliable and valid measurements in multiple studies (with some variations between them) [8], [9], there is no guaranty or method to evaluate that in absence of ground truth.

Therefore, there is a need for single measurement metric that represents the reliability of the entire study in the absence of ground truth. In this paper, we compare different inter-rater reliability metrics in terms of their sensitivity in identifying common issues associate to crowdsourcing studies.For this, we simulate a ground truth of user ratings following the discrete QNormal distribution according to the subject bias theory [10]. We investigate the impact of the the users' bias and rating uncertainty, as well as the presence of unreliable raters, on traditional reliability metrics. Finally, we identify a concrete guideline when the reliability of a crowdsourced QoE study is highly questionable, such that researchers are encouraged to revisit their study design in order to eventually improve the reliability of results from crowdsourcing-based quality studies. This means, for example, to cure issues stemming from the presence of unreliable users by implementing additional consistency checks and filtering mechanisms, or to cure issues stemming from high uncertainty among raters by providing more accurate instructions and better training of participants.

The paper is organized as follows. In Section II, we briefly review common inter-rater reliability metrics. The simulation process for obtaining ground truth rating data is explained

in Sections III and IV. The results of our analyses and the comparison of inter-rater reliability metrics are reported and discussed in Section V. Finally, a summary and proposals for future work conclude the paper in Section VI.

II. Reliability Metrics and Related Works

The concept of reliability relates to the overall consistency of a measure when assessing a trait [11]–[14]. This means that repetitions of a reliable experiment would provide essentially the same results. In contrast, the related concept of validity is the extent to which a measure actually measures the trait, and is out of scope of this work. Several metrics have been proposed in literature, which can be used to assess the reliability of a QoE experiment in terms of the inter-rater reliability, which is manifested in the agreement or consistency between the ratings of different participants.

The most commonly used approach to assess the inter-rater reliability is by computing rank correlation coefficients between the users' ratings, such as Spearman's rho (ρ) or Kendall's tau (τ) [15]. These correlation coefficients express the pairwise similarity of different users' rank ordering of stimuli, which is obtained from their QoE ratings, e.g., [16]. In this work, the average correlation $\bar{\rho}$ or $\bar{\tau}$ over all pairs of raters is considered to assess the reliability of the QoE experiment. Intraclass correlation (ICC) is a family of parametric correlation coefficients [17], [18], which is closely related to ANOVA, and compares the variance between raters with the variance over all ratings. For this, the metric can target either absolute agreement or consistency. In this work, $ICC(3,1)$ will be investigated, which is applicable to assess consistency in QoE experiments as it is based on the single ratings of a fixed set of raters, which rate all stimuli.

A more strict metric, which is based on absolute categorical agreement is Cohen's kappa [19]. It considers the share of equal ratings among two raters, i.e., per cent agreement, but additionally accounts for the possibility of agreement by chance. An extension for multiple raters was proposed by Fleiss [20], such that Fleiss' kappa (κ) can be applied for QoE experiments with a large group of participants. Note that κ considers ratings on a categorical rating scale. We hypothesize that κ will not be a strong metric for QoE ratings, which are given on an ordinal rating scale. However, due to its popularity, we include it in the evaluation.

Krippendorff's alpha (α) is a large family of reliability coefficients, which embraces several other reliability coefficients, and thus, can handle various scenarios, such as number of observers, levels of measurement (e.g., categorical, ordinal), sample sizes, and presence or absence of missing data [13]. It accomplishes this by calculating disagreements instead of correcting percent-agreements, thereby avoiding some limitations of other metrics.

Finally, we consider the SOS parameter (a) [21], which reflects the level of rating diversity across users for the different scenarios by relating the MOS values and the standard deviation of a QoE study. It allows for a compact statistical summary of subjective user tests, and it supports checking the reliability of test result data sets as well as their comparability across different QoE studies.

For the specific use case of the reliability of crowdsourced QoE experiements, several other related works exist, which are briefly outlined. [16] used reliability metrics not only to assess inter-rated reliability, but also intra-rater reliability in order to detect unreliable users. [22] presented two metrics for crowdsourcing studies based on task difficulty and resulting errors, which can be used to detect unreliable subjects. Also many other mechanisms and approaches for identifying and rejecting unreliable user ratings were proposed, e.g., [2], [6], such as consistency tests, content questions, gold standards, attention tests, or proper monitoring of the user device. Note, however, that intra-rater reliability and detection of unreliable users is out of scope of this work. Furthermore, [10] investigated reliability in crowdsourced QoE studies from a theoretical perspective, and presented the subject bias theory to explain user ratings using random variables. As our simulation model is based on this theory, it will be shortly outlined below. [23] presented an analysis method for detecting inconsistent subjective data by checking for typical or atypical rating score distributions. Finally, we do not want to correct subjective ratings of participants (in contrast to [24] for bias removal), but rather evaluate the reliability of the entire study as a whole.

III. Model for User Rating Distributions

To account for many possible outcomes in the rating behavior, for any stimuli and any users, a user's rating $R \sim RV(\mu,\sigma)$ is modeled with a random variable, which has mean μ and standard deviation σ. Here, μ represents the central tendency of the user's experience, and σ quantifies the uncertainty of the user. The subject bias theory [10] proposes a normal distribution, $R \sim \mathcal{N}(\mu,\sigma)$, which is then limited to the range $[1;5]$ and discretized by rounding. The bounded and rounded continuous normal distribution leads to the so-called *QNormal* distribution with parameters μ and σ: $Q \sim \mathrm{QNorm}(\mu,\sigma) = [\mathcal{N}(\mu,\sigma)]_1^5$. Please note that, in general, the expected user rating $\mathrm{E}[Q] = m = \mu$ and the observed standard deviation $\mathrm{Std}[Q] = s = \sigma$ due to the effect of bounding and rounding.

As discussed above, the not fully controlled environment in crowdsourcing studies can lead to biases in the resulting ratings. Here, we will focus on a scenario with no bias and three bias scenarios, namely, positive bias, mixed bias, and fake users. In the no bias scenario, all users are unbiased and rate according to $Q \sim \mathrm{QNorm}(\mu,\sigma)$. The three bias scenarios differ from the no bias scenario in that a certain share of unbiased ratings are replaced by the ratings of biased users.

a) Positive bias: In many studies, we observe a shift of user ratings towards higher scores for different reasons, such as pleasing the employer [6], [25]. This may be taken into account by shifting the mean μ of the underlying normal distribution by a constant $\beta > 0$. Typically values for β are in the order between 0.25 and 0.75, see, for example, [2], [6], [26], [27]. The resulting distribution of a positively biased user rating is $Q \sim \mathrm{QNorm}(\mu+\beta,\sigma), \beta > 0$.

b) Mixed bias: In addition to users with a positive bias, there may also be users which are not able to properly consume the test contents, e.g., due to noisy environment [3], hidden influence factors like improper devices [2], or different cultural perception of aesthetics [25]. This may again be taken into account by shifting the mean μ as above, this time using a negative constant $\beta < 0$. In the mixed bias scenario, both positively and negatively biased users can appear. Note that we will not cover a scenario with only negatively biased users, as it is mostly analogue to the positive bias scenario.

c) Fake users: Finally, there may be some workers who are not understanding the test or who are not conducting the test properly. As described in [2], these users are denoted as *fake users* if the user rating is uncorrelated to the stimuli and purely randomly selected. Although it may be easy to identify and filter fake users, we consider in our analysis both the presence of those fake users, as well as filtered QoE results without ratings from fake users. We take the discrete uniform distribution over all five categories, i.e., $Q \sim \mathrm{U}(1,5), P(Q = i) = \frac{1}{5}$, as the rating distribution of fake users.

Figure 1 illustrates bias β and uncertainty σ in the user rating distribution Q. The left plot shows the effects of bias on a user with a true average user rating of $\mu = 4.5$, who has a high uncertainty ($\sigma = 1.0$). If there is no bias (orange distribution), the resulting average user rating is $m = 4.32$, and we clearly observe the spread of the user rating across the rating scores from 2 to 5 due to the high uncertainty ($s = 0.80$). Note once again here that, following the subject bias theory, there is a difference in the true average user rating μ and true uncertainty σ, and the resulting average user rating m and resulting standard deviation s due to bounding and rounding, respectively. In case of negative bias ($\beta = -0.5$, blue distribution), the user rating distribution is shifted towards lower QoE ratings ($m = 3.93$), while the positive bias of $\beta = 0.5$ (green distribution) shifts probability towards the right ($m = 4.63$). Moreover, it can be seen that the resulting standard deviation becomes smaller if the mean user rating moves closer to the edges, which is a consequence of the bounded rating interval, and will be discussed below.

The remaining plots in Figure 1 illustrate the effects of uncertainty. The middle figure shows the resulting user rating distributions for a true average user rating of $\mu = 2.25$ and various levels of uncertainty σ. It can be seen that a very low level of uncertainty $\sigma = 0.1$ (green distribution) does not spread the user rating far from the underlying μ. Thus, rounding causes the resulting user rating distribution to almost degrade to a fully deterministic distribution approaching $m = 2$ and $s = 0$. As uncertainty increases, the spread of the user rating increases, which also causes the resulting distribution to flatten and spread. Note that for $0.5 \leq \sigma \leq 1$, the resulting m and s stay close to the true μ and σ. Eventually, for larger uncertainty σ, the user rating distribution degenerates more and more towards a discrete uniform distribution.

The right figure shows the resulting standard deviation of the user rating distribution depending on the resulting mean and the degree of uncertainty σ. The shape of the curves is similar to the MOS-SOS plots [21], however, we are only considering the rating of a single user (and hence no MOS values across users). We visualize the theoretical minimum standard deviation $s_{\min}$ (dashed line) given a mean user rating m, which is $s_{\min} = \sqrt{m^*(1-m^*)}$ with $m^* = m - \lfloor m \rfloor$. Here, it can be seen that a low uncertainty of $\sigma = 0.25$ is still very close to the theoretical minimum. For increasing uncertainty, also the resulting standard deviation increases as the user rating distribution degenerates towards a discrete uniform distribution, which would result in the maximum $s = \sqrt{2} \approx 1.41$ for $m = 3$.

This effect can be nicely illustrated considering the example of the middle plot of $\mu = 2.25$. When following the vertical line at $m = 2.25$ in the right plot, it can be seen that the minimum possible s is 0.43. As observed in the middle plot, for σ above the minimum s, the resulting m and s stay close to μ and σ. However, such low s is not possible for smaller σ, which results in a shifted m closer to the nearest integer and greatly reduced s for the user rating distribution. As σ approaches 0, the resulting user rating distribution generates towards a deterministic rating distribution. On the other end, as σ approaches $s = \sqrt{2}$, we eventually see a shift of m towards 3, which is a consequence of the degeneration towards a discrete uniform distribution. Further increasing σ towards its maximum value 2 will cause W- or U-shaped distributions due to the truncation of the tails of the normal distribution at the bounds of the rating scale, which might not be considered realistic rating distributions for a single user and a single stimulus, and thus, are out of scope of this work.

Finally, the plot also shows that the resulting standard deviation is decreasing towards the edges of the rating scale as the rating distribution is confined by the resulting mean user rating. For example, a resulting mean user rating closer to 5 would require more and more probability mass to reside in rating category 5, which leads to a decreasing resulting standard deviation almost independent of the underlying true uncertainty σ.

IV. Simulation of QoE Experiments

We simulate QoE experiments, in which the bias, the uncertainty, and the true user ratings are known. This ground truth is required to understand the influence of bias and uncertainty on reliability metrics applied to crowdsourced QoE studies. In particular, the absolute values of the reliability metrics can be put into relation with bias, uncertainty, or ratio of fake users. As a result, guidelines can be derived when a crowdourced QoE study is considered not to be reliable.

The simulated QoE experiment consists of k test conditions (or stimuli), for which each stimulus x has a certain true MOS value M_x for $x = 1, \ldots, k$. We use perfectly designed and selected stimuli, which result in equidistant true MOS values $M_x \in [1, 5]$. Hence, $M_x = 1 + (x-1)\frac{5-1}{k-1}$ for $x = 1, \ldots, k$. For each stimulus x, every user $u = 1, \ldots, n$ has the true central tendency (or true average user rating) corresponding to the true MOS of that stimulus, i.e., $\mu_x = M_x$. Thus, according to the subject bias theory, for all stimuli x and users u, the

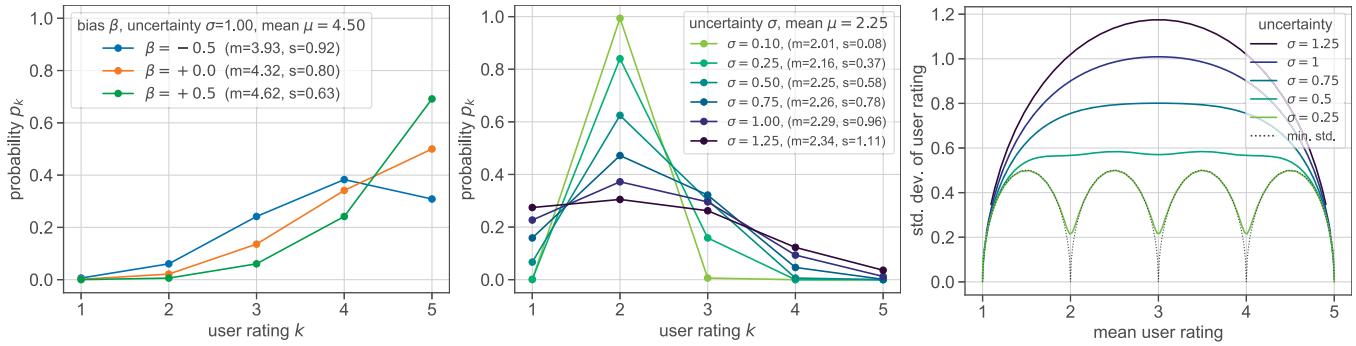

Fig. 1. **Illustration of Bias and Uncertainty.** Resulting user rating distributions for different bias β (left figure), different degrees of uncertainty σ (middle figure) as well as the first two moments, i.e. mean m and standard deviation s, of the resulting user rating distributions (right figure).

rating of a single user follows the QNormal distribution with parameters $\mu_x + \beta_u$ and σ_u due to the crowdsourcing setting:

$$Q_{u,x} \sim \mathrm{QNorm}(\mu_x + \beta_u, \sigma_u)$$

Please note that we are considering simplified scenarios, for which the bias is independent of the stimulus x. Nevertheless, this can be a realistic assumption, for example, in case users may please the employer (positive bias) or in case the degradation is masked by the effect of environment's condition. We are also assuming that the uncertainty of the users is not depending on the concrete stimulus, but rather on the type of service or application under test in the QoE study. This is inspired by the SOS hypothesis [21], where the user rating diversity is expressed by a single parameter only, which is denoted as the SOS parameter a in [21]. As we will see below, our simulation methodology also fulfills the SOS hypothesis, and there is a direct mapping between σ and a.

In a single QoE experiment, we are considering $k = 21$ stimuli or test conditions, and it follows that stimulus x has a true MOS value of $M_x = \frac{x+4}{5}$. Each stimulus is rated by $n = 30$ subjects. To investigate the impact of bias on reliability metrics, four scenarios are considered:

- *No bias*: The users are not biased, hence, we assume $\beta = 0$ for all users and all stimuli.
- *Positive bias*: Two different types of users are considered. One user class has no bias ($\beta = 0$), while the other group has a positive bias of $\beta = 0.5$. Each user is randomly assigned to one of the two classes with probability $\frac{1}{2}$.
- *Mixed bias*: Three different types of user classes are considered: no bias ($\beta = 0$), positive bias ($\beta = 0.5$), negative bias ($\beta = -0.5$). Each user is randomly assigned to one of the three user classes with probability $\frac{1}{3}$.
- *Fake users*: In this scenario, we do not want to simply replace regular ratings with the ratings of fake users, to keep the same amount of $n = 30$ regular users for all scenarios. Thus, we consider 15 additional fake users, such that the number of raters increases to $n_F = 45$. This results in a fake user ratio of $1/3$, which may be realistic in crowdsourced QoE studies, e.g., [6].

Another key influence factor on reliability metrics is the uncertainty of users. Therefore, we systematically investigate uncertainty and conduct a parameter sensitivity study on σ. In the parameter sensitivity study, we assume that the uncertainty of users depends on the service or application under test. Hence, we assume that all users have the same degree of uncertainty. Please note that the uncertainty of users may generally be higher in a crowdsourced setting due to remote test instructions compared to a laboratory study with a dedicated test moderator.

For each combination of bias scenario and uncertainty parameter, we simulated 1000 QoE experiments. In the following, we report average result scores of the simulated QoE experiments and the investigated reliability metrics. Note that, due to the high number of simulation runs, the confidence intervals are very small, such that they are omitted from the plots for better readability. The maximum confidence interval width is 0.0023 across all metrics and scenarios.

V. Discussion of Numerical Results

We analyze the inter-rater reliability of the users using the above introduced standard metrics, namely, average Spearman correlation $\bar{\rho}$, average Kendall correlation $\bar{\tau}$, intraclass correlation $ICC(3,1)$, Krippendorff's α, Fleiss' κ, as well as the SOS parameter a. The scope of this study is to describe the characteristics and behavior of these metrics for the simulated QoE experiments. In particular, we are interested to understand what is captured by the metrics and how sensitive they are with respect to the input parameters. These include bias, uncertainty, and fake users, as well as consistency across users. The ultimate goal is to provide thresholds when a QoE study can be considered to be reliable.

Figure 2 shows the main results of our analyses. Each of the six subplot represents a single metric's behavior in the different bias scenarios and for different uncertainty. In the following, we will elaborate on the main findings.

a) Bias: Comparing the investigated bias scenarios, namely, no bias (blue), positive bias (green), and mixed bias (orange), it can be seen from the overlapping curves that $\bar{\rho}$, $\bar{\tau}$, and $ICC(3,1)$ do not reveal the bias, since these metrics are quantifying the consistency of the user ratings, but do not consider the absolute agreement. Only κ, α, and a have the discriminative power to identify groups of biased users, since

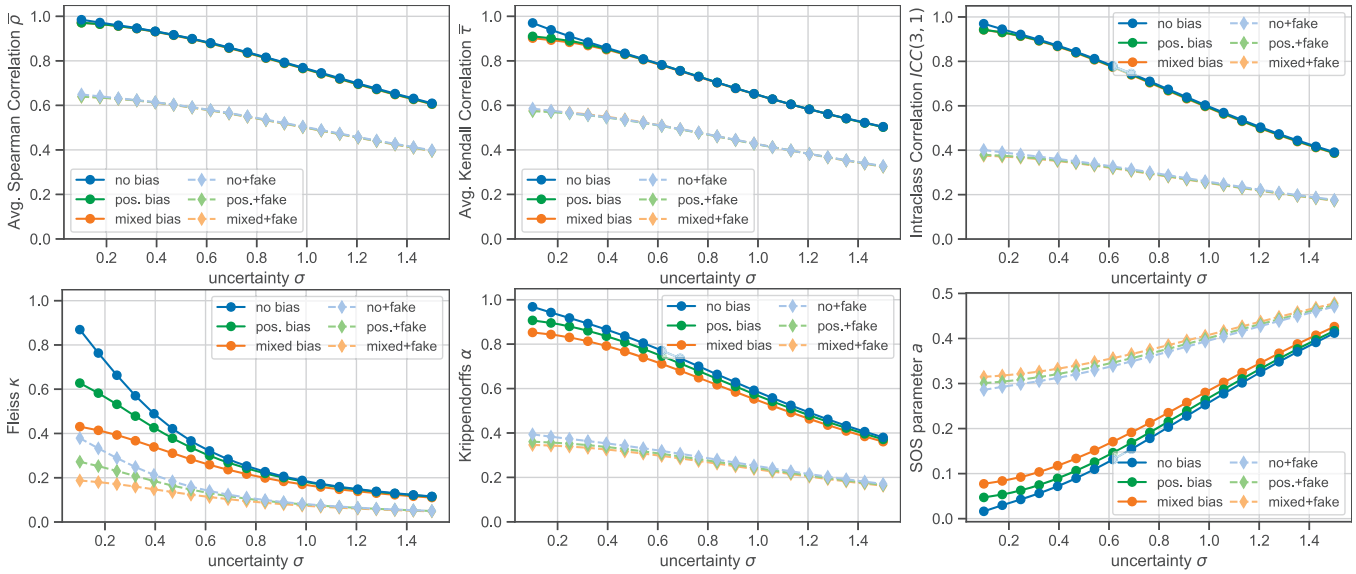

Fig. 2. **Parameter sensitivity study on uncertainty σ.** Comparison across metrics for fixed σ and different scenarios.

the corresponding curves are not overlapping. However, for α and a, the metrics' absolute differences are very low, and thus, bias could easily be confused with a lower uncertainty. Also for κ, the bias scenarios can only be well distinguished for a very small uncertainty of users. In fact, literature suggests $\kappa > 0.61$ [28] for substantial agreement, which we can only obtain in a real QoE experiment if we have almost deterministic user ratings, i.e., very confident users. Thus, the practical usage of all analyzed metrics to identify bias in crowdsourced QoE studies is rather limited.

b) Uncertainty: Regarding uncertainty, all subplots of Figure 2 confirm that the metrics are able to capture the underlying uncertainty. We generally observe a negative trend, such that the metrics' values decrease for increasing uncertainty, except for a, which follows an inverse trend. The $ICC(3,1)$, κ, α, and a are more sensitive than other metrics as they utilize wider range of the metric $[0;1]$.

c) Fake Users: When comparing the metrics for scenarios with (pale colors) and without fake users (strong colors), Figure 2 shows that $\bar{\rho}$, $\bar{\tau}$, $ICC(3,1)$, α, and a nicely separate the scenarios with and without fake users. The best separation can be achieved by $ICC(3,1)$ and α, which also avoid overlapping metric values even for high uncertainties. Again, $\bar{\rho}$ is equally suited, but struggles from not fully utilizing the metric range, while $\bar{\tau}$ and a already show overlapping values for smaller uncertainties.

Comparing the behavior of the reliability metrics, the correlations between all metrics are very high. Especially, $\bar{\rho}$, $\bar{\tau}$, $ICC(3,1)$, and α reach correlations above 0.98, while a shows negative correlations with absolute values above 0.96. This means that they capture similar properties of a QoE experiment. Only Fleiss kappa shows lower correlations ranging from 0.84 to 0.91, as it was designed for agreement on nominal data, and thus, cannot capture consistent trends in the ratings of participants on an ordinal rating scale.

d) Groups of Users: Beyond the parameter sensitivity study on the uncertainty σ, we consider now different user groups in terms of bias *and* uncertainty. Users are randomly assigned an uncertainty level with $\sigma \in \{0.25, 0.5, 0.75, 1., 1.25\}$ as well as a bias $\beta \in \{-0.5, 0, +0.5\}$. This scenario is referred to as *mixed bias + groups*. For comparing the uncertainty groups, we additionally investigate single uncertainty levels: $\sigma = 0.25$ (min.), 0.75 (avg.), and 1.25 (max.).

The numerical results indicate that the analysis of the uncertainty groups is not necessary. In fact, the scenario with the average uncertainty (i.e., average over all uncertainty values of the various groups) leads to almost the same absolute values of the reliability metrics. Hence, the parameter sensitivity study from above allows to derive conclusions with uncertainty groups. Hence, our limitation of fixed uncertainty in the results above in Figure 2 may be extended to user groups.

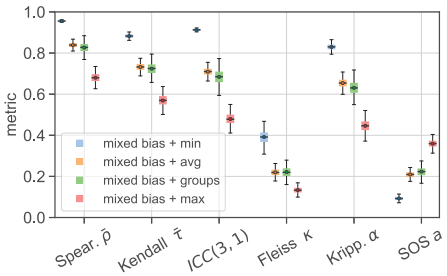

Fig. 3. **User Groups:** Mixed bias for different kind of uncertainity scenarios including mixed groups of uncertainty.

To sum up, when considering the performance in all investigated scenarios, $\bar{\rho}$, $\bar{\tau}$, and $ICC(3,1)$ performed well with respect to all criteria. They are unaffected by bias, and are able to detect both high uncertainty, as well as the presence of fake users. Moreover, as indicated by the very

high correlations, they all captured the same trends. Among these three, $ICC(3,1)$ shows a slight advantage by utilizing a wider range of values, however, $ICC(3,1)$ is limited to experiments, in which all participants rate all stimuli. Note that other ICC models exist for other experiment designs, in which all/a subset of stimuli is rated by all/a subset of participants, but these models were out of scope of this work.

Thus, also taking the ease of utilizing the average Spearman correlation $\bar{\rho}$ in almost all statistical software packages into account, we recommend $\bar{\rho}$ as inter-rater reliability metric for QoE experiments in crowdsourcing. As a rule of thumb, we propose that a QoE experiment with $\bar{\rho} < 0.75$ should be revisited by researchers. Note that a low score does not invalidate the QoE results, however, the reliability of the QoE experiment might be severely affected by high uncertainty among participants or the presence of fake users.

VI. Conclusion and Outlook

In this work, we simulated ratings of different user groups and scenarios typical in crowdsourcing (different bias scenarios, uncertainties, and fake users) following the discrete QNormal distribution according to the subject bias theory [10]. The results showed that both fake users and uncertainty could be detected by standard reliability metrics, while the detection of bias might not be possible in a realistic scenario. In practice, $\bar{\rho}$ is sufficient to evaluate inter-rater reliability in terms of consistency, and we might even define thresholds, e.g., $\bar{\rho} < 0.75$, for detecting the presence of high uncertainty or fake users. Scoring low on this metric should be a strong incentive for test moderators to revisit their test design. To fight uncertainty, the test could be better explained, while reliability checks should be implemented or adapted to filter out fake users. This can help researchers to eventually improve the reliability of results from crowdsourcing-based quality studies.

The simulation was required to create a ground of user ratings in typical scenarios of QoE experiments in crowdsourcing. However, our result should be consider preliminary and more sophisticated simulation with less assumptions should follow which ultimately leads to more realistic rating data. Namely in future work, we will consider other distributions for the underlying user ratings, different user groups and probabilities for each group, as well as different QoE experiment designs, e.g., experiments, in which participants only rate a subset of all stimuli. Furthermore, we will extend our analysis to other rating scales, such as continuous rating scales, and corresponding metrics. Finally, we will apply our method on real crowd-sourced datasets with and without embedded quality control mechanism to demonstrate its usage.

References

[1] P. Le Callet, S. Möller, A. Perkis, *et al.*, "Qualinet white paper on definitions of quality of experience," *European network on quality of experience in multimedia systems and services (COST Action IC 1003)*, vol. 3, no. 2012, 2012.

[2] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2013.

[3] B. Naderi, S. Möller, and G. Mittag, "Speech quality assessment in crowdsourcing: Influence of environmental noise," *DAGA*, pp. 299–302, 2018.

[4] B. Naderi and S. Möller, "Application of just-noticeable difference in quality as environment suitability test for crowdsourcing speech quality assessment task," in *2020 QoMEX*, IEEE, 2020, pp. 1–6.

[5] ITU-T Rec. P.808, *Subjective evaluation of speech quality with a crowdsourcing approach*. Geneva: International Telecommunication Union, 2018.

[6] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, "Best practices and recommendations for crowdsourced qoe - lessons learned from the qualinet task force crowdsourcing," 2014.

[7] B. Naderi, T. Polzehl, I. Wechsung, F. Köster, and S. Möller, "Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm," in *Proc. Interspeech*, 2015.

[8] B. Naderi, T. Hossfeld, M. Hirth, F. Metzger, S. Möller, and R. Z. Jiménez, "Impact of the number of votes on the reliability and validity of subjective speech quality assessment in the crowdsourcing approach," in *2020 QoMEX*, IEEE, 2020.

[9] B. Naderi, R. Z. Jiménez, M. Hirth, S. Möller, F. Metzger, and T. Hoßfeld, "Towards speech quality assessment using a crowdsourcing approach: Evaluation of standardized methods," *Quality and User Experience*, vol. 6, no. 1, pp. 1–21, 2020.

[10] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2210–2224, 2015.

[11] J. P. Guilford, "Psychometric methods," 1954.

[12] J. J. Bartko and W. T. Carpenter, "On the methods and theory of reliability.," *Journal of Nervous and Mental Disease*, 1976.

[13] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Communication methods and measures*, vol. 1, no. 1, pp. 77–89, 2007.

[14] K. A. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," *Tutorials in quantitative methods for psychology*, vol. 8, no. 1, p. 23, 2012.

[15] M. G. Kendall, "Rank correlation methods.," 1948.

[16] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of youtube qoe via crowdsourcing," in *2011 IEEE International Symposium on Multimedia*, IEEE, 2011, pp. 494–499.

[17] J. J. Bartko, "On various intraclass correlation reliability coefficients.," *Psychological bulletin*, vol. 83, no. 5, p. 762, 1976.

[18] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability.," *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979.

[19] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[20] J. L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[21] T. Hossfeld, R. Schatz, and S. Egger, "Sos: The mos is not enough!" In *2011 QoMEX*, IEEE, 2011, pp. 131–136.

[22] L. Janowski, "Task-based subject validation: Reliability metrics," in *2012 Fourth International Workshop on Quality of Multimedia Experience*, IEEE, 2012, pp. 182–187.

[23] J. Nawala, L. Janowski, B. Cmiel, and K. Rusek, "Describing subjective experiment consistency by p-value p–p plot," in *28th ACM International Conference on Multimedia*, Seattle, WA, USA, 2020.

[24] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *2017 Data compression conference (DCC)*, IEEE, 2017, pp. 52–61.

[25] M. Varela, T. Mäki, L. Skorin-Kapov, and T. Hoßfeld, "Increasing payments in crowdsourcing: Don't look a gift horse in the mouth," in *4th international workshop on perceptual quality of systems (PQS 2013). Vienna, Austria*, 2013.

[26] T. Volk, C. Keimel, M. Moosmeier, and K. Diepold, "Crowdsourcing vs. laboratory experiments–qoe evaluation of binaural playback in a teleconference scenario," *Computer Networks*, vol. 90, pp. 99–109, 2015.

[27] T. Hoßfeld and C. Keimel, "Crowdsourcing in qoe evaluation," in *Quality of experience*, Springer, 2014, pp. 315–327.

[28] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.